

RESEARCH ARTICLE

Credit Risk Assessment Using Alternative Financial and Behavioral Data

Arif Zaman¹ ✉, Md Mehedi Hassan², Mst Mannujan Akther³, Md Kabir Chowdhury⁴, Sanjida Islam⁵ and Abdullah Hill Hussain⁶

¹ National University

² Washington University of Science and Technology (WUST)

³ General Software Solutions

⁴ Gannon University

⁵ Bangladesh National University

⁶ University of Liberal Arts Bangladesh

Other author E-mails: Md Mehedi Hassan — Mehedi61@gmail.com; Mst Mannujan Akther — falguni.mst@gmail.com; Md Kabir Chowdhury — mdkabirchowdhury05@gmail.com; Sanjida Islam — Sanjidaislam0910@gmail.com; Abdullah Hill Hussain — nirzhorshovon@gmail.com
ORCIDs: Arif Zaman — 0009-0004-3159-4785; Md Mehedi Hassan — 0009-0001-8910-0846; Mst Mannujan Akther — 0009-0008-5428-8747; Md Kabir Chowdhury — 0009-0009-3563-6926; Abdullah Hill Hussain — 0009-0008-0436-027X

Corresponding Author: Arif Zaman, **E-mail:** arifzamman286@gmail.com

ABSTRACT

Credit scoring draws on transaction records, bill payments, and digitally observed behavior to evaluate applicants whose conventional credit files are sparse. Whether these signals improve lending decisions without creating new forms of exclusion remains unresolved. This study synthesizes pre-2023 research on alternative financial and behavioral data and develops a transparent, simulation-based comparison of conventional and augmented credit-risk models. Thirty-five peer-reviewed journal articles inform the synthesis; a separately identified conference proceeding provides context on small-business finance in Bangladesh. The demonstration generates 24,000 synthetic borrower records with six conventional predictors, eight consent-dependent alternative indicators, and an explicitly specified default mechanism. Four classifiers are evaluated on held-out cases using discrimination, precision-recall, calibration-sensitive loss, and operational screening measures. In the simulation, augmenting logistic regression increases receiver operating characteristic area from 0.740 to 0.810 and average precision from 0.387 to 0.521; random forests and gradient boosting achieve areas of 0.796 and 0.803, respectively. Shapley-based explanations identify bill-payment regularity, inflow volatility, and debt burden as prominent simulated predictors. These figures describe a controlled demonstration, not observed improvements in real lending portfolios, because the outcome mechanism deliberately contains behavioral information. The reviewed field evidence nonetheless indicates that properly governed alternative data can complement formal credit histories, particularly for thin-file applicants. A viable implementation therefore requires temporally valid measurement, consent and minimization, external validation, probability calibration, group-level fairness audits, and accessible adverse-action explanations. The contribution is an evaluation design that separates predictive feasibility from credible claims of financial inclusion, while making the assumptions behind the numerical illustration explicit.

KEYWORDS

Credit scoring; alternative data; behavioral finance; financial inclusion; model interpretability; algorithmic fairness; synthetic benchmarking

ARTICLE INFORMATION

ACCEPTED: 11 November 2022

PUBLISHED: 25 December 2022

DOI: 10.32996/jefas.2022.4.4.22

1. Introduction

Credit allocation is an information problem before it is an algorithmic problem. Lenders must estimate the likelihood that borrowers will miss contractual payments, yet the information available at application differs sharply across households, firms, and countries. Traditional scorecards summarize income, existing obligations, prior delinquencies, and bureau histories. These measures work best when applicants have already participated in formal credit markets. For a new borrower, the absence of a bureau record can be mistaken for evidence of excessive risk. Hand and Henley (1997) and Crook et al. (2007) established the statistical foundations of credit classification while recognizing that observed applicant samples and measurement choices constrain conclusions.

Alternative data seek to narrow that information gap. Account cash flows, recurring bill settlement, utility payment histories, and transaction consistency may reveal capacity and willingness to repay before a long borrowing history develops. A different group of signals comes from digital footprints, such as mobile usage or website interactions. The distinction matters. Transaction volatility has a direct financial interpretation; a device or browsing characteristic may correlate with default but be difficult to justify in a consequential decision. Khandani et al. (2010) demonstrated the usefulness of transactional information combined with bureau data. Berg et al. (2020) found that online digital footprints contained information about consumer default and complemented conventional scores. Neither result establishes that every available digital feature should be collected.

The social significance extends beyond predictive accuracy. Financial constraints hinder smaller enterprises more severely than larger firms (Beck et al., 2005). Hossain and Hassan (2019), discussing cooperative investment and small-business development in Bangladesh, provide a relevant development context rather than direct evidence for a particular scoring algorithm. Access to credit depends simultaneously on information, lending rules, local institutions, and the cost of capital. An apparently better score will not necessarily widen access if a lender uses it to exclude more marginal applicants or sets prices that remain unaffordable. Jagtiani and Lemieux (2018, 2019) show why lending penetration, grading, and risk pricing should be examined together.

A further distinction concerns the moment of prediction. Application scoring estimates risk before granting a new loan; behavioral scoring updates the outlook for an existing account. Payment streams recorded after origination cannot legitimately serve as predictors of the original approval decision. Mixing these settings can produce impressive but misleading performance. It also complicates comparisons between lenders that hold rich account histories and newcomers that observe only the applicant's consented external records. The relevant benchmark is always the information actually available for the specific decision, at the time that decision must be made.

Three questions organize this paper. First, which alternative financial and behavioral indicators have demonstrated incremental predictive value, and in what populations? Second, how should models be evaluated when default events are unevenly distributed, probabilities matter more than labels, and protected groups may experience different error rates? Third, what should count as sufficient evidence before a lender relies on behavioral signals to deny, price, or renew a loan? These questions connect statistical performance with the institutional conditions under which a model becomes useful.

The article combines a critical synthesis of journal literature published no later than December 2022 with a fully disclosed simulation. Simulation makes it possible to compare feature sets and inspect receiver operating characteristic curves and Shapley explanations without misrepresenting inaccessible borrower data. It also imposes a fundamental boundary: an outcome deliberately generated partly from alternative variables will favor models capable of observing those variables. Numerical results are therefore methodological illustrations, not empirical tests of inclusion. The practical contribution is a validation framework that identifies the evidence still needed before reported gains can support real credit decisions.

2. Literature Review

Conventional credit scoring evolved from discriminant analysis and logistic regression toward increasingly flexible classification systems. Hand and Henley (1997) documented both the statistical choices available to lenders and the unusual sampling problems created when default is only observed after credit has been granted. Baensens et al. (2003) compared competing classification algorithms, and Crook et al. (2007) extended the discussion to account management throughout a credit relationship. Subsequent benchmarks found substantial variation across datasets rather than one universally superior method (Lessmann et al., 2015). Louzada et al. (2016) traced the field's methodological diversification, while Dastile et al. (2020) noted continuing challenges involving interpretability and imbalance. A recurring lesson is that a modeling innovation should be compared against a carefully tuned, transparent baseline rather than against an artificially weak conventional score.

Evidence for account behavior as a credit signal predates the current interest in digital platforms. Khandani et al. (2010) combined commercial-bank transaction information and bureau records to forecast consumer delinquency. Their study is important because transactions observe cash-management behavior closer to repayment than many fixed demographic attributes do. Yeh and Lien (2009) likewise examined the predictive value of credit-card payment histories and demonstrated why probabilities of default deserve attention beyond binary class labels. Serrano-Cinca et al. (2015) found that platform grades, indebtedness, and borrower characteristics helped explain defaults in peer-to-peer loans. Emekter et al. (2015) examined credit risk and loan performance on comparable platforms. These studies describe observed financial behavior, although institutional underwriting and borrower selection still influence which repayment outcomes can be seen.

Online credit markets widened the range of potentially useful information. Lin et al. (2013) reported that borrowers' friendship networks on a lending platform were associated with funding, pricing, and ex post performance. Iyer et al. (2016) showed that aggregated lender judgments using soft information could predict outcomes that were not fully summarized by visible credit categories. Malekipirbazari and Aksakalli (2015) used random forests for social-lending risk assessment. These findings are not interchangeable. Platform friendships may reflect endorsement or social capital; lender judgments may incorporate narrative signals; automated ensemble classifiers may simply model financial attributes more flexibly. Their predictive success depends on the platform's institutions and should not be translated into an unrestricted invitation to profile social relationships.

The most direct alternative-data studies examine digitally recorded behavior. Óskarsdóttir et al. (2019) used mobile calling patterns and network structure, reporting gains from phone-derived information and differences between statistical and profit-based criteria. Björkegren and Grissen (2020) matched mobile-phone usage with credit repayment, finding predictive information among individuals with limited formal histories. Berg et al. (2020) analyzed simple website-access and registration traces and found complementarity with credit bureau scores. Djeundje et al. (2021) evaluated email usage and psychometric measures, reporting improvement over models based only on demographic attributes while stressing the sensitivity of performance to data splits. Together these studies support selective information value, not an assertion that a particular app, communication partner, or psychological response is causally related to creditworthiness.

Financial inclusion is an additional, separate empirical outcome. Beck et al. (2005) found disproportionate growth constraints among smaller firms, motivating interest in lower-cost screening. Jagtiani and Lemieux (2018) reported that fintech lending reached some markets less served by traditional banks. Their later analysis found that platform loan grades and conventional bureau scores were not perfectly aligned and that alternative assessment could affect loan classification and pricing (Jagtiani & Lemieux, 2019). Fuster et al. (2019) documented operational efficiency in technology-enabled mortgage lending. None of these studies implies automatic welfare improvement for every applicant. Expanded origination can coexist with elevated repayment stress, higher rates, or restricted access for groups whose digital histories are less extensive.

Algorithm selection changes both statistical error and governance requirements. Brown and Mues (2012) found that forests and boosting handled pronounced default imbalance comparatively well. Xia et al. (2017) examined optimized boosted trees, and Papouškova and Hajek (2019) connected default forecasting to expected-loss modeling. Addo et al. (2018) assessed machine- and deep-learning approaches; Lessmann et al. (2015) demonstrated the value of evaluating multiple criteria. Yet model complexity creates a potential explanation deficit. Bussmann et al. (2020) proposed explainable credit-risk modeling approaches, and Lundberg et al. (2020) provided methods for interpreting tree predictions. Rudin (2019) cautioned that post hoc explanations are not equivalent to the transparency of a directly interpretable model, especially where decisions have serious consequences.

Discrimination and stability make accuracy an insufficient endpoint. Barocas and Selbst (2016) explain how apparently neutral input data can encode social disadvantage. Chouldechova (2017) showed tensions among common fairness criteria when outcome prevalence differs across groups. Fuster et al. (2022) found unequal distributional effects from machine-learning mortgage prediction, while Bartlett et al. (2022) documented persistent disparities in mortgage pricing. These studies do not establish that every alternative variable is discriminatory. They establish why historical outcomes, group composition, and decision thresholds must be examined rather than treating sensitive attributes' removal as an adequate safeguard.

Credit-risk research also exposes a persistent measurement mismatch. Application-stage models are frequently tested on accounts whose outcomes are fully observed, whereas applicants who were denied credit have no comparable repayment records. Selection then shapes the observed correlation between each feature and default. A behavioral feature that appears harmless among established cardholders might behave differently among first-time borrowers. Conversely, signals extracted from a digital service could largely reproduce platform participation rather than underlying financial resilience.

Source legitimacy deserves its own place in the research map. Bank statements and utility receipts generally describe completed economic activities, although their collection can still expose intimate details. Mobile metadata and social connections are more ambiguous. The number of calls, digital contacts, or login events may be shaped by employment schedules, connectivity, household composition, and temporary disruptions. Predictive correlations do not settle whether such features are necessary for extending credit. The empirical record should therefore be read alongside data minimization and an explicit account of whose information is collected, who can challenge errors, and whether a less intrusive substitute performs similarly. A model that improves a score but cannot explain a denial imposes costs that published discrimination statistics alone will not capture.

The studies reviewed here also differ in prediction horizon and business objective. A probability useful for a thirty-day servicing intervention may be poorly suited to a twelve-month origination decision. More complete comparisons would harmonize the default definition, observation window, feature timestamp, funding population, and cost of error. Without that alignment, contrasting published accuracy percentages risks treating noncomparable outcomes as if they were direct contests between algorithms.

Finally, evaluation metrics must match lending decisions. Hand (2009) questioned uncritical reliance on the area under the receiver operating characteristic curve. Saito and Rehmsmeier (2015) explained why precision-recall analysis is informative when events are relatively rare. Van Calster et al. (2019) emphasized probability calibration; a high-ranking model can still systematically overstate or understate risk. Comparing only headline accuracy obscures missed defaults, unnecessary denials, and differences in exposure. The literature therefore motivates a layered assessment of discrimination, calibrated probability, decision cost, temporal transportability, and fairness. Table 1 organizes representative findings by data source and principal caution. The synthesis highlights a gap between proof that a signal contains information and proof that collecting and using it improves decisions ethically and economically. That gap defines the methodological contribution developed next.

Table 1. Selected journal evidence on alternative signals, predictive value, and limitations.

Study	Signal / setting	Principal evidence	Interpretive caution
Khandani et al. (2010)	Transactions + bureau	Machine-learning models using bank records improved delinquency forecasting	Single-bank context; restricted access
Lin et al. (2013); Iyer et al. (2016)	Social ties + soft information	Platform networks and lender judgments contained risk information	Platform selection and social proxy concerns
Óskarsdóttir et al. (2019)	Mobile call behavior	Calling-pattern features added prediction and profit-oriented signals	Connectivity and relationship privacy
Björkegren & Grissen (2020)	Mobile metadata	Phone-use signatures predicted repayment among thin-file customers	Population and telecom specificity
Berg et al. (2020)	Digital footprint	Website footprints complemented bureau scores	Proxy discrimination; consent
Djeundje et al. (2021)	Email + psychometrics	Nontraditional attributes improved demographic baselines	Split sensitivity; collection ethics
Jagtiani & Lemieux (2019)	Platform grades + loan data	Alternative underwriting related to pricing and loan performance	Not a causal access estimate
Fuster et al. (2022); Bartlett et al. (2022)	Mortgage decisions	Technology can produce uneven borrower outcomes	Group-level lending consequences

Note. Findings summarize distinct studies; effect estimates are not pooled because samples, outcomes, and evaluation protocols differ.

3. Methodology

This study adopts a critical narrative synthesis with a reproducible, simulation-based benchmarking component. Relevant peer-reviewed journal articles published through December 2022 were identified through topic-specific searches covering credit scoring, cash-flow underwriting, mobile behavioral signals, peer-to-peer lending, class imbalance, explainability, and algorithmic fairness. Priority was given to original empirical studies, substantial methodological benchmarks, and foundational governance research. The review is purposive rather than a registered systematic review or quantitative meta-analysis; no pooled effect size is estimated. Thirty-five journal articles provide the main scholarly base. A separate 2019 conference proceeding supplied by the author is cited for contextual discussion of small-business finance, without being treated as empirical validation of alternative scoring.

Because lender-level transaction and mobile records are generally proprietary and were not accessible for this work, the numerical comparison uses entirely synthetic records. The simulation contains 24,000 borrowers generated with NumPy's random generator and seed 2022. Its binary target indicates a simulated default event over a stylized credit horizon. The final sample contains 3,818 generated defaults, a prevalence of 15.91 percent. No individual in the data represents a real applicant, and the model outputs cannot be interpreted as field estimates. Rather than conceal this constraint, the data-generating rules deliberately make it inspectable.

Six conventional predictors describe age, logarithmic income, debt-to-income ratio, previous delinquency count, loan term in months, and loan-to-income ratio. Eight augmented variables represent inflow volatility, the share of days with a positive account balance, income regularity, the share of bills paid on time, mobile top-up regularity, utility-payment lag count, a monthly cash buffer index, and app-usage stability. These indicators are synthetic approximations of consent-dependent features discussed in the literature; they are not direct reproductions of observed bank feeds or telecommunications metadata. Sex, race, religion, and geographic identifiers are intentionally excluded from the model inputs, although their exclusion alone would not guarantee fairness in real applications. Table 2 supplies operational descriptions and governance qualifications for the feature groups.

The simulator first draws a common latent financial-strain variable and uses it to introduce correlation among debt, delinquency, liquidity, and routine stability. Continuous financial variables arise from clipped normal, beta, or lognormal distributions, while delinquency and utility-lag indicators are generated from count distributions. Default probability follows a logistic generating equation with additive conventional and behavioral terms plus a small interaction between high debt burden and payment irregularity. Coefficients increase simulated risk for prior delinquency, debt burden, volatile inflows, and utility delays and decrease it for stronger buffers, regular receipts, and timely bill settlement. A Bernoulli draw converts each probability into a binary label. These modeling choices encode the hypothesis under study; consequently, any improvement from adding the encoded features is expected and does not independently confirm their value outside the simulation.

Records were divided using stratification into 14,400 training cases, 4,800 validation cases, and 4,800 independent test cases. Random state 42 fixed the split and all model seeds. A future real-data study should replace the random split with a forward-time separation based on application date, preserve lender and cohort boundaries, and prohibit any feature recorded after the decision cutoff. The synthetic sample lacks dates and cannot evaluate this risk. In practice, statements, phone records, and bill-payment aggregates must be constructed strictly from data already available when the application is decided. Administrative labels such as later collection outcomes are not permissible predictive inputs.

Four supervised classifications were fitted. The reference specification uses standardized conventional predictors in regularized logistic regression. The augmented logistic specification retains identical estimation settings and adds eight alternative signals, isolating information-set differences. A random forest with 180 trees, minimum leaf size 16, and square-root feature sampling tests nonlinear relationships. Gradient boosting uses 170 shallow trees of maximum depth three, learning rate 0.045, subsampling 0.85, column sampling 0.90, minimum child weight 15, and regularization parameter four. The two tree ensembles use the augmented features. Parameters were fixed before the reported test evaluation, rather than being tuned on test outcomes. The design supports an incremental-information comparison for logistic regression and a separate comparison of augmented algorithm families; it does not identify the causal impact of any single data provider.

Predicted probabilities were evaluated using receiver operating characteristic area (ROC-AUC), average precision, and Brier score. ROC-AUC measures relative ranking over all thresholds. Average precision summarizes precision-recall performance and is sensitive to event prevalence. Brier score penalizes squared deviation between assigned probability and realized binary outcome and therefore evaluates a probability-sensitive loss, although it does not replace inspection of calibration plots. In addition, each model's validation predictions defined a screening cutoff at the eightieth percentile. Applying that frozen cutoff to the test observations yields recall and precision for an approximately top-fifth risk-review queue. This is an illustration of review capacity, not an optimal lending threshold. Actual accept-or-decline decisions require explicit loss-given-default, exposure, return, and affordability assumptions.

Figure 1 specifies the conceptual workflow from source assessment to accountable decisions. Figure 2 displays test-set ROC curves, drawn from fitted simulation models rather than illustrative hand sketches. Figure 3 uses exact tree-based Shapley contributions from a subset of 350 held-out synthetic cases, aggregated as mean absolute feature contributions in log-odds units. SHAP describes a fitted prediction, not a causal mechanism: correlated alternatives may divide attribution unpredictably, and a strong attribution may still be inappropriate for use in credit decisions. Explanations should therefore be paired with feature governance and counterfactual review rather than interpreted as automatic evidence of fairness.

The conventional benchmark intentionally retains age, which appears in many historical underwriting datasets, so that the demonstration resembles published scorecard comparisons. That inclusion is not an endorsement of age-based decision rules in any jurisdiction. The eight alternative indicators would also require a source register documenting consent scope, provenance, refresh frequency, aggregation window, missing-data mechanism, and the precise decision for which each feature may be used.

A proper real-world augmentation analysis should distinguish additional information from broader population coverage. Reporting only results among customers able and willing to share data would give a selective picture of inclusion. Missing alternative information must not automatically be interpreted as greater credit risk, and a conventional scoring path should remain available when possible.

Robustness analyses were deliberately bounded in this demonstration. The models were not exposed to synthetic downturns, corrupted transaction histories, adversarial manipulation, differing borrower cohorts, or cross-country transfers. Those stress tests would be necessary to examine whether a predictor captures relatively durable capacity or merely transient behavior. Confidence intervals were also not presented as population uncertainty estimates, because sampling uncertainty around a simulated mechanism would leave the larger uncertainty about that mechanism untouched. The tabulated point estimates permit direct verification of the computational exercise but should not be generalized beyond its assumptions.

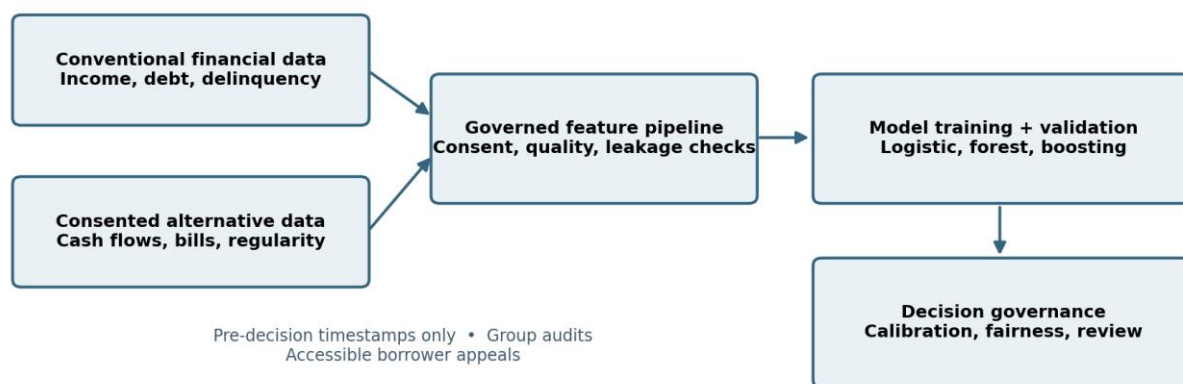
Replicability relies on the disclosed generation design, fixed random states, unambiguous feature partitions, and reported hyperparameters. In a lender deployment, additional tests would include population stability, calibration intercept and slope, missingness by group, subgroup false-positive and false-negative rates, adverse-action reason fidelity, and performance by first-time versus established borrowers. Model monitoring should use institutionally justified intervention thresholds and a documented appeals process. Ethical approval was not required for the simulation because it used no human participant records. Any extension involving account, telecom, or psychometric information would require lawful collection, meaningful consent, retention limits, and review of disparate impact before operational use.

Table 2. Variable families and operational meanings in the synthetic benchmark.

Group	Features	Construct	Governance issue
Conventional	Age; log income	Applicant profile and logged earnings	Consider legality and fairness
Conventional	Debt-to-income; loan-to-income	Repayment pressure and requested exposure	Use decision-time values
Conventional	Prior delinquencies; loan term	Past arrears count and term in months	Ensure predecision cutoff
Alternative financial	Inflow volatility	Dispersion of incoming funds	Consent; avoid postdecision data
Alternative financial	Positive balance days share	Proportion of time above zero balance	Protect transaction confidentiality
Alternative financial	Income regularity	Consistency of receipt intervals	Allow irregular employment patterns
Alternative financial	On-time bill share	Share of simulated bills settled promptly	Check disputed or missing bills
Alternative behavioral	Mobile top-up regularity	Consistency of recharge behavior	Avoid penalizing limited access
Alternative financial	Utility payment lag	Count of delayed utility settlements	Verify source and billing cycles
Alternative financial	Monthly cash buffer	Synthetic liquidity reserve indicator	Respect emergency spending
Alternative behavioral	App-usage stability	Stability of consented usage activity	High sensitivity; consider exclusion

Note. The complete design contains six conventional and eight augmented predictors; grouped rows describe related fields. All values are simulated, not scraped or collected from people.

Figure 1. Information governance and model evaluation framework.



Note. Original conceptual framework; source eligibility, temporal validity, model evaluation, and borrower recourse are separate checks.

4. Discussion

The simulation illustrates incremental information without establishing a field effect. Logistic regression trained on conventional variables reached ROC-AUC 0.740 and average precision 0.387. When eight generated behavioral and alternative financial measures were included, the same algorithm reached ROC-AUC 0.810 and average precision 0.521. Its Brier score declined from 0.118 to 0.104. These changes are substantial in this simulation, but the generator itself assigned outcome information to payment regularity, cash-buffer measures, and inflow volatility. The exercise cannot determine how informative those variables would be for a real lender or a population not represented by formal banking records. Table 3 presents all four test-set comparisons with that qualification made explicit.

The ranking of algorithms is revealing for a different reason. The augmented random forest produced ROC-AUC 0.796, while gradient boosting reached 0.803, both slightly below augmented logistic regression. Their average precision values were 0.500 and 0.505, respectively. This finding does not overturn work reporting strong ensemble performance in credit data (Brown & Mues, 2012; Lessmann et al., 2015; Xia et al., 2017). Instead, it reflects a simulation whose default function is predominantly logistic and additive. Flexible classifiers are not entitled to win when the underlying relationships favor a simpler model. Conversely, an

observed improvement from boosted trees in real data might reflect interactions, complex missingness patterns, or data leakage; benchmark superiority should never be assumed before population-specific validation.

Decision thresholds change the meaning of model quality. In a review queue defined by the top fifth of validation risk scores, conventional logistic regression identified 44.4 percent of defaults in the test sample with 37.2 percent precision. The augmented logistic specification identified 55.4 percent with 44.0 percent precision. These are screening statistics, not realized profits. A lender might use such a queue to request missing documents, conduct manual review, or offer counseling. Translating it directly into automatic rejection would demand a stronger justification: the cost of unnecessarily denying credit may vary substantially between a solvent thin-file borrower and an applicant already facing overindebtedness.

Feature explanations illuminate where the modeled information originates, not whether the collection of those features is legitimate. The gradient-boosting SHAP summary assigns the largest mean absolute contribution to on-time bill share, followed by inflow volatility, debt-to-income ratio, monthly cash buffer, and earlier delinquencies. This ordering is consistent with the generator's coefficients and with the behavioral interpretation of the variables. It does not establish that mobile application patterns capture personal reliability, nor that a high contribution indicates a causal link. A borrower experiencing temporary income disruption could generate patterns similar to a structurally high-risk borrower. Explanations should present both uncertainty and available alternatives to unsupported automated decisions (Bussmann et al., 2020; Lundberg et al., 2020; Rudin, 2019).

The field literature provides stronger, but still conditional, reasons to investigate alternative data. Björkegren and Grissen (2020) found that mobile usage was predictive of repayment within their setting, including among people with limited financial histories. Berg et al. (2020) showed that basic digital footprints complemented bureau information. Djeundje et al. (2021) reported useful email and psychometric signals but also notable sensitivity to sample partitions. Óskarsdóttir et al. (2019) observed that statistical ranking and profit-based performance could favor different combinations of telephone-network features. The relevant conclusion is that information may exist outside bureaus. The unsupported leap would be to claim that mobile logs, social graphs, and repayment cash flows are equally reliable or ethically interchangeable.

A practical ranking of data sources should begin with proximity to the lending obligation. Consented bank-account inflows, verified recurring income, overdraft frequency, rent and utility payment continuity, and shortfall episodes have relatively direct financial interpretations. Digital communication networks, typing patterns, device metadata, and psychometric questions may have weaker causal explanations and greater potential to encode protected characteristics. Their collection also expands the consequences of a privacy breach. Strict minimization matters even when a provider offers a technically useful variable. The fair-lending question is not only whether the feature is legal to possess, but whether it is proportionate, necessary, contestable, and materially beneficial over less intrusive alternatives.

The distribution of gains merits close scrutiny. Alternative scoring could increase access for households with no bureau file while worsening terms for applicants whose phone use is irregular because of work, disability, age, rural coverage, or device sharing. Some of those dimensions can correlate with protected status even if the explicit attribute is removed. Barocas and Selbst (2016) explain the mechanisms through which proxies reproduce disadvantage; Fuster et al. (2022) and Bartlett et al. (2022) show why distributional questions cannot be settled by overall prediction. Chouldechova (2017) also demonstrates that different fairness criteria may conflict when group prevalence differs. A responsible analysis should report multiple measures, state which tradeoffs are being accepted, and avoid disguising normative decisions as purely technical optimization.

Historical underwriting creates another identification problem. Default labels are observed only for applicants who obtained credit under existing rules. If a new score identifies many previously rejected borrowers as apparently safe, their future performance remains unknown. Training exclusively on approved accounts can therefore reward the old policy, even when augmented variables are informative. A carefully designed pilot with controlled exposure, documentation of selection mechanisms, and subsequent repayment follow-up would be more persuasive than a retrospective score comparison. The study should separate the impact of changing data coverage from the impact of changing the model or the institution's appetite for risk. That separation is especially important when making claims about inclusion.

Economic evaluation similarly requires more than ROC curves. A probability can rank borrowers well yet systematically overpredict losses in a new cohort. Van Calster et al. (2019) emphasize calibration, and Hand (2009) warns against reducing decision quality to a single threshold-free number. Operational validation should present calibration by decile, precision-recall analysis under observed prevalence, profit and expected-loss simulations under prespecified assumptions, and sensitivity to macroeconomic shifts. For small loans, the incremental cost of collecting and verifying consented alternative information may outweigh a small statistical gain. For larger exposures, even a modest improvement may matter if it is stable and accompanied by sufficiently informative adverse-action reasons.

The institutional setting also shapes whether inclusion is a plausible outcome. Beck et al. (2005) identify structural financial obstacles faced by smaller firms, while Hossain and Hassan (2019) address cooperative investment as one response to small-business financing constraints in Bangladesh. This motivates, but cannot replace, local validation. Jagtiani and Lemieux (2018,

2019) show that fintech distribution and pricing may differ from traditional channels; this may depend as much on business design and competition as on predictive inputs. For borrowers, meaningful inclusion requires affordable access, understandable terms, and manageable repayment, rather than merely an additional route to algorithmic evaluation.

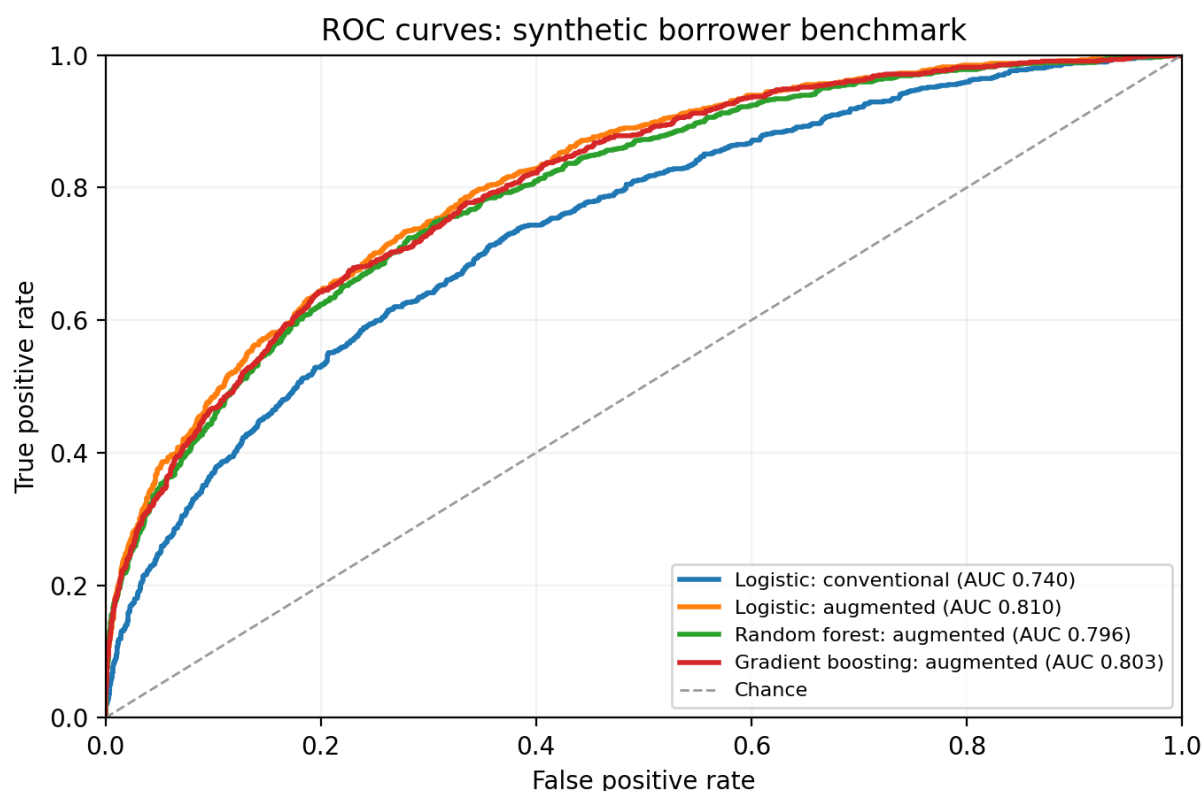
Overall, the evidence favors a restrained interpretation. Alternative financial records can be valuable especially when conventional files omit regular payment behavior. Some digital traces contain predictive information, but their benefits are context dependent and their governance burden can be considerable. The simulation demonstrates a useful comparison architecture, not a universal uplift estimate. Before deployment, researchers must replace synthetic assumptions with observed, consented, time-stamped data, benchmark against strong baselines, calibrate probabilities, audit distributional effects, and verify that changed decisions actually benefit the intended borrowers. That is a more demanding standard than maximizing accuracy, but it matches the real stakes of allocating credit.

Table 3. Held-out model comparison from the synthetic borrower simulation.

Specification	ROC-AUC	Avg. precision	Brier ↓	Recall ¹	Precision ¹
Logistic: conventional	0.740	0.387	0.118	44.4%	37.2%
Logistic: augmented	0.810	0.521	0.104	55.4%	44.0%
Random forest: augmented	0.796	0.500	0.108	52.9%	42.1%
Gradient boosting: augmented	0.803	0.505	0.106	53.0%	42.5%

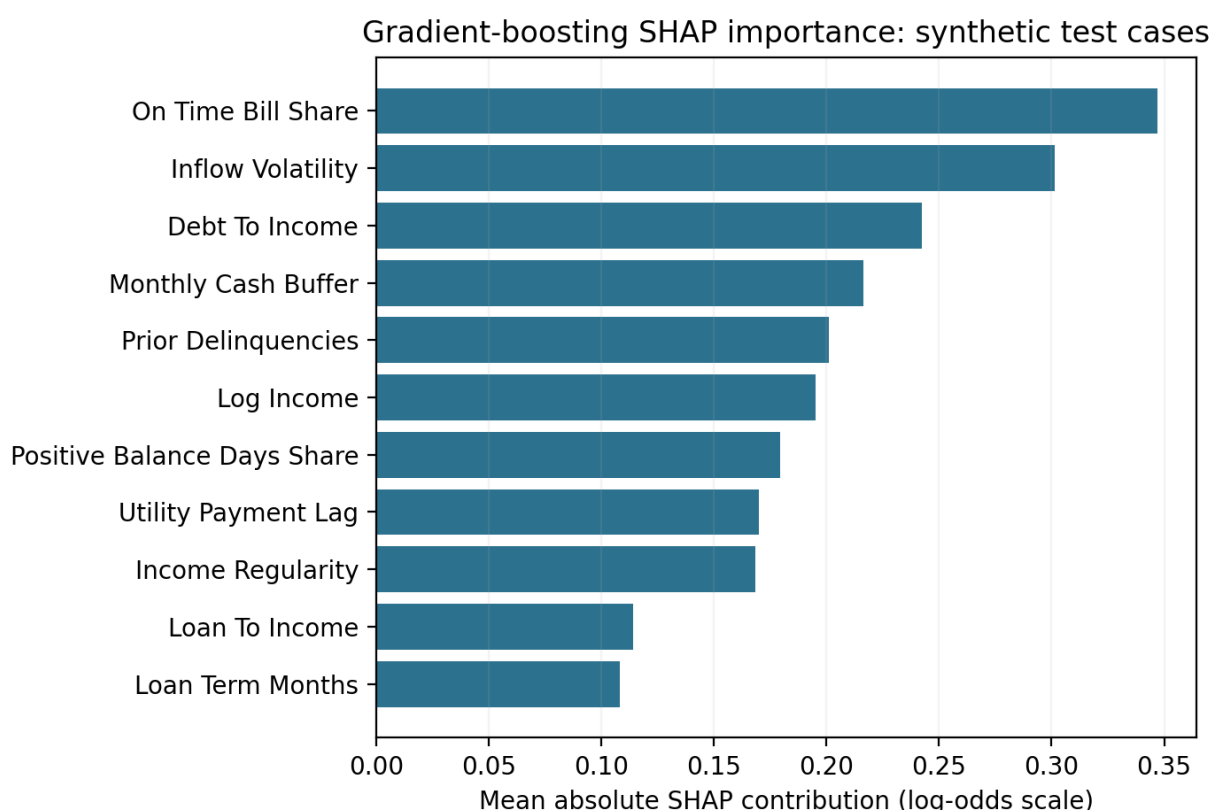
Note. $n = 4,800$ test records; 15.91% overall simulated default prevalence. ¹ Top-fifth risk-review cutoff fixed on validation data. Figures are synthetic benchmarks, not estimates from live borrowers.

Figure 2. Receiver operating characteristic curves on the synthetic test set.



Note. ROC curves and AUC values were computed using fitted synthetic models. They do not represent any lender portfolio.

Figure 3. Global SHAP feature importance for augmented gradient boosting.



Note. Mean absolute TreeSHAP values for 350 synthetic held-out cases; contributions are predictive explanations on the log-odds scale, not causal effects.

5. Conclusion

Alternative financial and behavioral data may improve credit-risk assessment by revealing repayment-relevant patterns that conventional files overlook. The pre-2023 literature provides evidence that bank transactions, bill payments, mobile-phone usage, and digital footprints can contain additional predictive information, for applicants with limited histories. Yet information value is not the same as permission to collect a feature or evidence that using it will increase financial inclusion.

This paper combines that literature with a disclosed synthetic benchmark. Adding alternative indicators raised logistic regression ROC-AUC from 0.740 to 0.810 in the simulated test sample, while tree ensembles did not outperform the augmented linear model. These numbers are properties of the assumed generating mechanism, not estimates from actual borrowers. Their purpose is to demonstrate a transparent feature-ablation and evaluation workflow.

For practice, the priority should be consented, financially interpretable measures backed by reliable timestamps and clear borrower recourse. Model comparisons should assess probability calibration, default detection, decision costs, and disparities across relevant groups, not accuracy alone. Future field studies must test whether alternative information changes approvals and pricing without imposing disproportionate harm. Only then can an improved risk score be interpreted as a contribution to responsible lending rather than simply a more elaborate screening instrument.

6. Limitations and Future Directions

The central limitation is reliance on synthetic borrower records. The outcome generator explicitly incorporates alternative indicators, so improved augmented-model performance is partly built into the experiment. Simulated feature distributions, correlations, event prevalence, and measurement quality do not reproduce any verified lending population. The random split cannot establish forward-looking stability under economic shocks, regulatory changes, or changes in borrowing behavior. SHAP contributions describe simulated fitted predictions and should not be read as causal financial mechanisms.

The literature synthesis is purposive rather than exhaustive, includes studies with heterogeneous outcomes and populations, and does not estimate a pooled effect. Proprietary data, rejected applicants' unobserved repayment outcomes, and varying lender selection rules limit comparability. The requested Bangladesh cooperative-finance proceeding supports institutional context, not a claim about alternative credit-scoring performance.

The next research step is a preregistered, multi-institutional evaluation with consented transaction and bill-payment histories, application-time feature snapshots, and a forward-time holdout. Designs should compare bureau-only, cash-flow-only, combined,

and privacy-minimized specifications. Reports should include calibration, uncertainty intervals, expected loss, approval and affordability outcomes, subgroup diagnostics, and temporal drift. Controlled lending pilots could assess people with previously missing scores while limiting borrower harm and documenting selection. Independent audits, deletion policies, access controls, and challenge procedures would make improved prediction more defensible. Future work should also evaluate whether simple, directly interpretable models capture most incremental information before adopting intrusive data sources or opaque architectures.

Future evaluations should distinguish technical explainability from borrower comprehension and test whether reason codes actually support correction of disputed information or meaningful appeals.

Funding: Funding information was not supplied in the manuscript.

Conflicts of Interest: A conflict-of-interest declaration was not supplied in the manuscript.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

- Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep learning models. *Risks*, 6(2), 38. <https://doi.org/10.3390/risks6020038>
- Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6), 627–635. <https://doi.org/10.1057/palgrave.jors.2601545>
- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3), 671–732. <https://doi.org/10.15779/Z38BG31>
- Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics*, 143(1), 30–56. <https://doi.org/10.1016/j.jfineco.2021.05.047>
- Beck, T., Demirgüç-Kunt, A., & Maksimovic, V. (2005). Financial and legal constraints to growth: Does firm size matter? *The Journal of Finance*, 60(1), 137–177. <https://doi.org/10.1111/j.1540-6261.2005.00727.x>
- Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of FinTechs: Credit scoring using digital footprints. *The Review of Financial Studies*, 33(7), 2845–2897. <https://doi.org/10.1093/rfs/hhz099>
- Björkegren, D., & Grissen, D. (2020). Behavior revealed in mobile phone usage predicts credit repayment. *The World Bank Economic Review*, 34(3), 618–634. <https://doi.org/10.1093/wber/lhz006>
- Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3), 3446–3453. <https://doi.org/10.1016/j.eswa.2011.09.033>
- Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in FinTech risk management. *Frontiers in Artificial Intelligence*, 3, Article 26. <https://doi.org/10.3389/frai.2020.00026>
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153–163. <https://doi.org/10.1089/big.2016.0047>
- Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3), 1447–1465. <https://doi.org/10.1016/j.ejor.2006.09.100>
- Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91, Article 106263. <https://doi.org/10.1016/j.asoc.2020.106263>
- Djeundje, V. B., Crook, J., Calabrese, R., & Hamid, M. (2021). Enhancing credit scoring with alternative data. *Expert Systems with Applications*, 163, Article 113766. <https://doi.org/10.1016/j.eswa.2020.113766>
- Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online peer-to-peer (P2P) lending. *Applied Economics*, 47(1), 54–70. <https://doi.org/10.1080/00036846.2014.962222>
- Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance*, 77(1), 5–47. <https://doi.org/10.1111/jofi.13090>
- Fuster, A., Plosser, M., Schnabl, P., & Vickery, J. (2019). The role of technology in mortgage lending. *The Review of Financial Studies*, 32(5), 1854–1899. <https://academic.oup.com/rfs/article/32/5/1854/5427795>
- Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1), 103–123. <https://doi.org/10.1007/s10994-009-5119-5>
- Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3), 523–541. <https://doi.org/10.1111/j.1467-985X.1997.00078.x>

- Hossain, M. N., & Hassan, M. M. (2019). Co-operative investment policy for growth and development of small and medium size businesses and improve the poverty condition in Bangladesh. In Proceedings of the 7th Mediterranean Interdisciplinary Forum on Social Sciences and Humanities (MIFS 2019) (pp. 142–154).
- Iyer, R., Khwaja, A. I., Luttmer, E. F. P., & Shue, K. (2016). Screening peers softly: Inferring the quality of small borrowers. *Management Science*, 62(6), 1554–1577. <https://doi.org/10.1287/mnsc.2015.2181>
- Jagtiani, J., & Lemieux, C. (2018). Do fintech lenders penetrate areas that are underserved by traditional banks? *Journal of Economics and Business*, 100, 43–54. <https://doi.org/10.1016/j.jeconbus.2018.03.001>
- Jagtiani, J., & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. *Financial Management*, 48(4), 1009–1029. <https://doi.org/10.1111/fima.12295>
- Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11), 2767–2787. <https://doi.org/10.1016/j.jbankfin.2010.06.001>
- Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136. <https://doi.org/10.1016/j.ejor.2015.05.030>
- Lin, M., Prabhala, N. R., & Viswanathan, S. (2013). Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science*, 59(1), 17–35. <https://doi.org/10.1287/mnsc.1120.1560>
- Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, 21(2), 117–134. <https://doi.org/10.1016/j.sorms.2016.10.001>
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2, 56–67. <https://doi.org/10.1038/s42256-019-0138-9>
- Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42(10), 4621–4631. <https://doi.org/10.1016/j.eswa.2015.02.001>
- Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing*, 74, 26–39. <https://doi.org/10.1016/j.asoc.2018.10.004>
- Papoukova, M., & Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision Support Systems*, 118, 33–45. <https://doi.org/10.1016/j.dss.2019.01.002>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432. <https://doi.org/10.1371/journal.pone.0118432>
- Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in P2P lending. *PLOS ONE*, 10(10), e0139427. <https://doi.org/10.1371/journal.pone.0139427>
- Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, 17, Article 230. <https://doi.org/10.1186/s12916-019-1466-7>
- Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, 78, 225–241. <https://doi.org/10.1016/j.eswa.2017.02.017>
- Yeh, I.-C., & Lien, C.-H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2), 2473–2480. <https://doi.org/10.1016/j.eswa.2007.12.020>